Universitat de Lleida

Estrategias de identificación de variantes asociadas a caracteres productivos mediante chips o secuenciación a baja cobertura

TRABAJO FINAL DE MÁSTER

Alumna: Inés Samperi Tena

Curso: 2019/2020

Tutor: Roger Ros Freixedes

Cotutor: Joan Estany Illa

Grupo: Departamento Ciencia Animal

Máster Ingeniería Agronómica

Índice

Resumen

Los chips de genotipado y la secuenciación permiten determinar el genotipo de un individuo para un elevado número de variantes a lo largo del genoma de forma automática. La información proveniente de estas tecnologías permite identificar variantes asociadas a un carácter de interés productivo y realizar predicciones genómicas del valor de mejora en especies ganaderas. La secuenciación permite identificar y genotipar variantes no incluidas en los chips de genotipado y conseguir así información más completa del genoma de un animal. Sin embargo, aunque su coste está disminuyendo, la secuenciación es una tecnología más cara que los chips de genotipado y hay una necesidad de encontrar estrategias para optimizar el uso de los datos de secuenciación, por ejemplo, usando coberturas moderadas o bajas (<7x). En este trabajo (i) se realizó un estudio de asociación genómica, (ii) se estudió la capacidad de detección de variantes y la precisión de los genotipos que nos ofrece la secuenciación a coberturas bajas o moderadas, y (iii) se exploraron qué estrategias de genotipado combinando chips de genotipado, secuenciación y técnicas de imputación permitían genotipar una población con la mayor precisión con un coste económico determinado. Se usaron los datos recogidos por chips de genotipado sobre 395 cerdos Duroc, con el fin de identificar regiones genómicas ligadas al pH de la carne, mediante un estudio de asociación genómica. Proponemos el gen *ATP1A1* como gen candidato asociado al pH en carne. En cuanto a la capacidad de detección de variantes mediante secuenciación a coberturas bajas y moderadas, se identificaron más del 95% de variantes presentes en el chip con las coberturas ≥5x, se disminuyó al 92% a una cobertura de 2x y a 83% a 1x. Por otro lado, se obtuvieron concordancias genotípicas y alélicas superiores al 80% con las coberturas ≥5x, a 2x el porcentaje disminuyó al 58% y para 1x al 35%. A nivel individual las coberturas 7x o 5x mostraron ser muy interesantes al permitir conseguir buenas concordancias genéticas y reducir el coste de secuenciación al compararlas con coberturas superiores. Sin embargo, se concluyó que a nivel poblacional y cuando hay técnicas de imputación disponibles, la cobertura de 2x es la mejor estrategia de secuenciación porque permite obtener datos sobre más individuos. La precisión de imputación conseguida fue mayor al aumentar los animales genotipados por medio de chips y la densidad del chip de genotipado no afectó la precisión de la imputación. En caso de presupuestos limitados (<10 €/animal) es aconsejable invertir más en chips, para poder realizar una buena imputación, y solo cuando los presupuestos para el genotipado son elevados se puede invertir en secuenciación.

Abstract

Genotyping chips and sequencing allow the genotype of an individual to be determined automatically for a high number of variants throughout the genome. The information from these technologies allows the identification of variants associated with productive traits and the genomic prediction of breeding values in livestock. Sequencing allows variants not included in genotyping chips to be identified and genotyped, thus obtaining more complete information on the genome of an animal. However, although the cost of sequencing is decreasing, sequencing is more expensive than genotyping chips and there is a need to find strategies to reduce the cost of using sequencing data, for example using moderate or low coverages (<7x). In this work (i) we carried out a genome-wide association study, (ii) we studied the variant discovery rate and the accuracy of the genotype calls offered by sequencing at moderate or low coverages, and (iii) we explored which genotyping strategies that combine genotyping chips, sequencing and imputation techniques provided the highest genotype accuracy with a determined economic cost. The data collected by genotyping chips on 395 Duroc pigs were used in order to identify genomic regions linked to meat pH, through a genome-wide association study. We propose the *ATP1A1* gene as a candidate gene associated with pH in meat. Regarding the variant discovery rate of sequencing at low and moderate coverage, more than 95% of the variants present on the chip were identified with the ≥5x coverage, but this rate decreased to 92% at 2x coverage and to 83% to 1x. On the other hand, genotypic and allelic concordances higher than 80% were obtained with coverage ≥5x, at 2x the percentage decreased to 58% and for 1x to 35%. At the individual level, the 7x or 5x coverage proved to be very interesting, allowing good genetic concordance to be achieved and reducing the cost of sequencing when compared to superior coverage. However, it was concluded that at the population level and when imputation techniques are available, 2x coverage is the best sequencing strategy because it allows for the sequencing of a larger number of individuals. The imputation accuracy achieved increased with increasing number of animals genotyped with the chip, and the density of the genotyping chip did not affect the precision of the imputation. In case of limited budgets (<10 €/animal), it is better to invest more in genotyping chips, and investment in sequencing should be considered only when budgets for genotyping are large.

1. INTRODUCCIÓN

1.1 Técnicas genómicas en mejora animal

La mejora animal se lleva realizando desde los comienzos de la domesticación de animales. Al principio, la selección se hacía observando las características del animal de forma visual, fijándose únicamente en su fenotipo (fruto de los genes y el ambiente). El desarrollo de la genética cuantitativa permitió estimar el componente genético heredable correspondiente a los caracteres productivos y que los animales se pudieran seleccionar por su mérito genético y transmitirlo a la siguiente generación.

El avance de la ciencia ha permitido desarrollar la genómica, la cual consiste en el estudio, cuantificación e identificación de los genomas. La selección genómica permite encontrar el máximo número de variaciones heredables presentes en el genoma de forma precisa. De esta forma se pueden usar los genotipos de cada animal para seleccionar aquellos que tengan un valor genético superior y así mejorar las poblaciones (Cañón, 2006). La selección genómica ha permitido aumentar la precisión de estimación de los valores genéticos y la intensidad de selección, a la vez que disminuir el intervalo generacional (Ángel-Martín et al., 2013; Fujita, 2007).

En la era actual se dispone de herramientas biotecnológicas que permiten analizar las variaciones presentes a lo largo de un genoma y conocer el genotipo de cada animal en las posiciones polimórficas de forma eficiente y a un coste asequible. De entre estas, las principales son los chips de genotipado y la secuenciación masiva. En poblaciones con genotipos conocidos se pueden identificar variantes asociadas a caracteres productivos o realizar predicciones del valor genético basadas en el genoma.

1.1.1 Estudios de asociación

Los estudios de asociación GWAS (en inglés, *Genome-Wide Association Studies*) permiten identificar cuáles son las regiones o los genes que se encuentran ligados a un carácter. Por medio de las tecnologías de genotipado (chips de genotipado o secuenciación) se determinan los genotipos con el fin de testar estadísticamente su asociación con un carácter de interés. Posteriormente se examinan cuáles son las variantes más asociadas y su posición en el genoma para tratar de identificar la variante o gen candidato causante de la variación observada. No se requiere de una hipótesis previa de asociación entre un gen y el carácter a estudiar (Ros-Freixedes et al., 2016; Estany & Pena, 2017).

Existen fenotipos que dependen únicamente de un gen, pero la mayoría están afectados por más de un gen (caracteres complejos). Entre variantes genéticas puede existir desequilibrio de ligamiento, es decir, que no segreguen de manera independiente. Esto es más probable que ocurra cuando los genes se encuentran próximos (en un mismo cromosoma) y que así se hereden juntos (Iniesta et al., 2005). Los GWAS se basan en el desequilibrio de ligamiento entre marcadores y en particular entre el que existe entre ellos y las variantes causales del efecto que se encuentran alrededor suyo. Se conoce

como marcadores genéticos o moleculares a las variaciones en un segmento o posición del ADN.

Se han descrito muchas variantes asociadas a algún carácter de interés, pero pocas han sido comprobadas que tengan un efecto causal, y todavía queda mucha varianza que no se explica por medio de estas asociaciones (CM Dekkers, 2012).

1.1.2 Predicción genómica

La predicción genómica consiste en predecir el valor de mejora de un animal a partir del conjunto de sus genotipos. Tradicionalmente, los valores de mejora se predecían usando metodologías como el BLUP *(Best linear unbiased prediction)* y teniendo en cuenta la información genealógica. La predicción genómica permite estimar el efecto de cada variante sobre el carácter de interés usando una población de referencia con fenotipos conocidos y calcular un valor de mejora genómico a partir de la suma de los efectos de cada una de las variantes consideradas. Esta técnica permite aumentar la precisión de los valores de mejora y puede permitir una mayor precisión e intensidad de la selección. Además, como solo requiere conocer los genotipos, la predicción genómica permite estimar los fenotipos a edades tempranas, adelantar el momento de selección y, por lo tanto, acortar significativamente el intervalo generacional (Van Eenennaam et al., 2014).

1.2 Tecnologías de genotipado

La precisión de la predicción y los GWAS dependen de las tecnologías de genotipado que se utilicen. Los chips de genotipado y la secuenciación del genoma son algunas de las tecnologías más utilizadas en el ámbito de la genómica. Estas tecnologías se basan en que la posición de los genes dentro del genoma se mantiene en todos los individuos de una misma especie, lo que permite encontrar variaciones entre diferentes animales.

Hay diferentes tipos de marcadores, pero nos centraremos en los SNP (polimorfismos de un solo nucleótido), que son el tipo de variación más usado hoy en día (Ángel-Martín et al., 2013; Fujita, 2007).

1.2.1 Chips de genotipado

Los chips de genotipado han sido y siguen siendo muy utilizados en la actualidad para conocer el genotipo de los individuos a estudiar. Los chips consisten en pequeñas secuencias de ADN (sondas) que se colocan de una manera ordenada sobre una superficie solida (normalmente de sílice). Estas sondas son diferentes fragmentos de cadenas complementarias a la cadena de ADN donde se encuentran las variantes a estudiar. Los pasos para poder obtener resultados por medio de chips de genotipado se observan en la Figura 1.

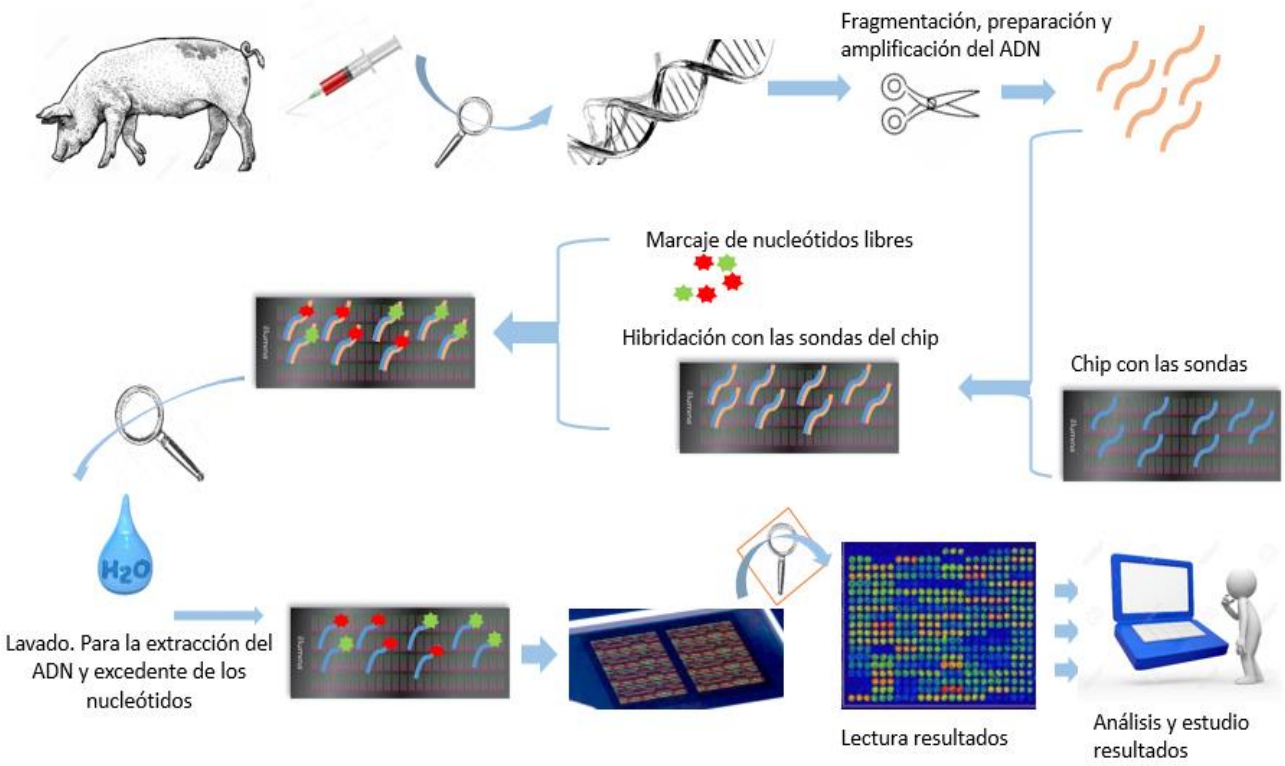


Figura 1. Proceso de genotipado mediante chips. 1) Extracción de la muestra de ADN a estudiar. 2) Fragmentación, preparación y amplificación del ADN. 3) Hibridación de los fragmentos de ADN con las sondas previamente diseñadas en el chip. 4) Marcaje de los nucleótidos libres. 5) Hibridación de los nucleótidos marcados con el ADN presente en la sonda. 6) Lavado. 7) Lectura de los resultados de hibridación. 8) Cuantificación de la lectura. 9) Análisis y estudio de los resultados.

Esta tecnología trabaja únicamente con marcadores tipo SNP y permite conocer el genotipo solo para variantes en aquellas posiciones predeterminadas que se han incluido en los chips durante su diseño.

Los chips se diseñan para incluir una serie de marcadores dispuestos a lo largo del genoma de la especie. La densidad del chip influirá en los resultados de los GWAS y en la predicción genómica. Cuanto más denso sea un chip de genotipado, la distancia entre marcadores y las variantes causales se reduce, aumentando así la posibilidad de encontrar variantes asociadas a los caracteres que se quieren estudiar y la precisión de la asociación y consecuentemente mejorar la predicción del valor genómico (Busquets & Agustí, 2001; Estany & Pena, 2017).

Hoy en día hay diferentes empresas que comercializan chips, con distinto número y tipo de SNP. Los chips de genotipado comerciales disponibles actualmente para porcino se pueden observar en la Tabla 1.

Tabla 1: Chips comerciales. Datos extraídos de las páginas oficiales de GeneSeek, Illumina y Affymetrics.

Proveedor	Nombre comercial	Nº SNPs	Distancia media entre SNPs (Kb)
Illumina	GGP Porcine LD Array	>10.000	250
GeneSeek	GGP Porcine BeadChip	>51.000	43
Illumina	GGP Porcine HD Array	70.000	42
Affymetrics	Affymetrix Axiom Porcine Array	658,692	3.34

1.2.2 Secuenciación

Una tecnología novedosa y revolucionaria de genotipado es la secuenciación de nueva generación o secuenciación masiva. Esta herramienta tan potente permite estudiar el genoma completo de un individuo y, junto con los avances en el campo de la bioinformática, está cambiando el ámbito de aplicación de la genómica. La secuenciación facilita detectar muchas más variaciones que los chips de genotipado, lo que permite identificar nuevas variantes asociadas a la variabilidad genética, tanto en especies ganaderas como en cultivos (Daetwyler et al., 2014; Schaid et al., 2018 y Yano et al., 2016;).

Un secuenciador lee de forma ordenada los nucleótidos de los diferentes fragmentos de ADN del genoma de un animal. En el proceso de secuenciación, el ADN del animal se fracciona en pequeños fragmentos (en torno a 200-300 pares de bases). Se leen todos los nucleótidos de cada uno de los fragmentos. Las lecturas de las secuencias se pueden ensamblar para obtener el genoma completo o bien se pueden comparar con un genoma de referencia para encontrar aquellas posiciones variantes.

Las principales ventajas de la secuenciación masiva son que permite secuenciar gran cantidad de ADN en poco tiempo, que detecta SNP pero también otro tipo de variantes, como inserciones y deleciones (cortas como largas), y que es capaz de hacer aflorar todas las variantes de una población, incluidas aquellas que son específicas de cada población en particular y no solo aquellas predeterminadas en el chip. En definitiva, la secuenciación masiva facilita tener acceso a todos los polimorfismos causales y conlleva una identificación más eficiente de los genes asociados a los caracteres de interés (Das et al., 2015; Gudbjartsson et al., 2015; Mordoh, 2019).

Durante la última década se ha producido un gran progreso en las técnicas de secuenciación, de tal manera que se han logrado numerosos avances en el rendimiento (Figura 2), tanto en volumen de procesado, velocidad como longitud de lectura, y reducción de costes (Van Dijk, et al., 2014).

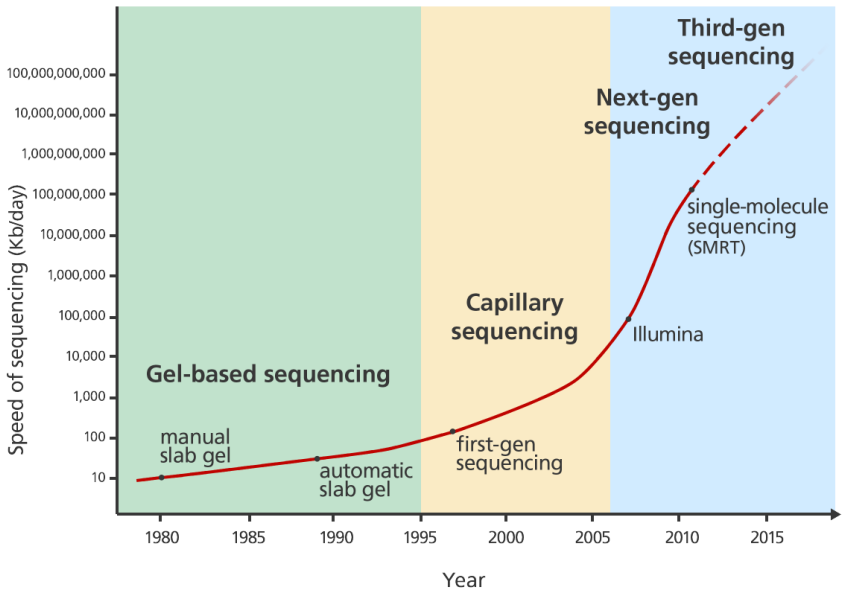


Figura 2. Evolución sobre las técnicas de la secuenciación. En las dos primeras regiones del gráfico se muestran las tecnologías basadas en los métodos de Sanger. Sobre la primera década del 2000 apareció la secuenciación masiva. Es importante observar cómo ha ido aumentado el rendimiento (línea roja). Fuente: https://www.yourgenome.org/stories/third-generation-sequencing

Actualmente los costes de los chips de genotipado y de la secuenciación están bajando y siguen bajando. Lo que permite que el uso de estas tecnologías sea día a día más viable económicamente a la escala necesaria para aprovechar los beneficios (Figura 3).

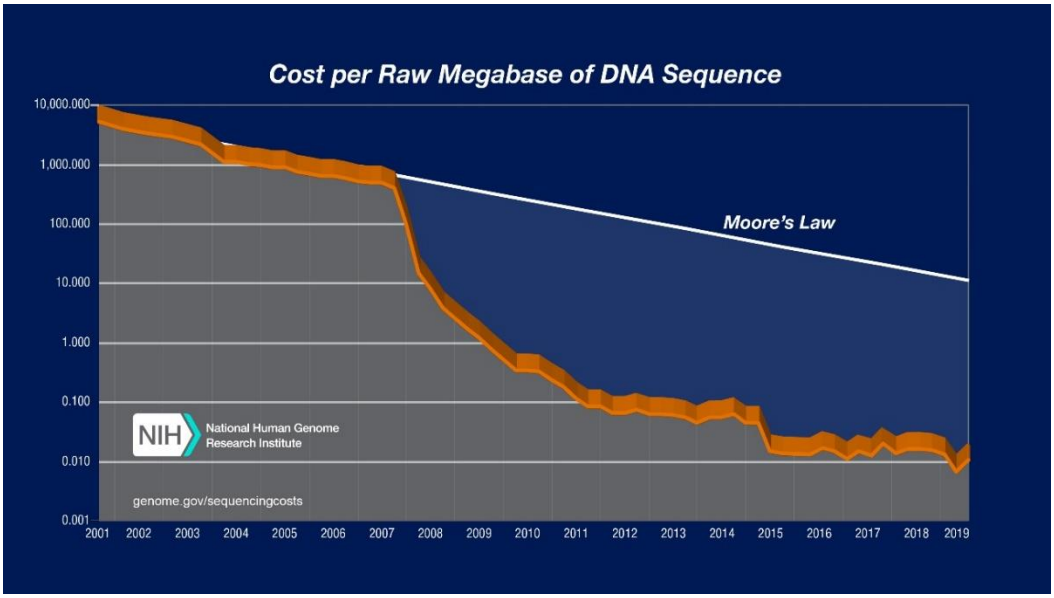


Figura 3. Reducción de costes en secuenciación por megabase a una calidad específica desde el año 2001 hasta la actualidad. La tendencia de la ley de Moore describe los avances a largo plazo, producidos en la industria del hardware y la informática sobre la potencia de los ordenadores. Los avances en el ámbito tecnológico se comparan con esta tendencia. La disminución por debajo de la ley de Moore indica que la capacidad de generación de datos de secuenciación es mayor que la capacidad de generación de equipos informáticos para almacenar y analizar dichos datos. Fuente: (genome.gov)

Sin embargo, el coste de secuenciar el genoma completo de un animal es muy variable y depende sobretodo del tamaño del genoma de la especie y de la cobertura de secuenciación (Mordoh, 2019; Ros-Freixedes et al., 2020b). La cobertura es el número promedio de veces que se lee cada una de las bases a lo largo del genoma (Mordoh, 2019). La cobertura de secuenciación determina la capacidad de encontrar variantes y la certidumbre con la que conocemos los genotipos correctos. La secuenciación a altas coberturas posee una gran fiabilidad, pero conlleva un elevado precio, mientras que cuando bajamos la cobertura reducimos el precio, pero a la vez se pierde fiabilidad de la información. Las estrategias de secuenciación de mayor interés son aquellas que permiten maximizar el número de individuos secuenciados a una cobertura concreta y con un coste total de secuenciación determinado. Con estas estrategias se obtiene un mayor índice de detección de variantes (es decir, se encuentran más variantes en el genoma del animal al compararlo con el genoma de referencia), sobre todo para variantes que se encuentran en una población con baja frecuencia (Ros-Freixedes et al., 2017)

Al realizar secuenciación a bajas coberturas se obtiene menos información sobre el genoma, por ejemplo, al obtener los resultados de secuenciación a una cobertura de 3x significa que de media entre todos los fragmentos de ADN estudiados habrá tres lecturas, es decir, en algunas partes de genoma se poseerán 3 lecturas, en otras 5, en otras 0, etc (Figura 4). En la Figura 5 se puede observar las limitaciones y posibles errores que se pueden cometer secuenciando a baja cobertura.

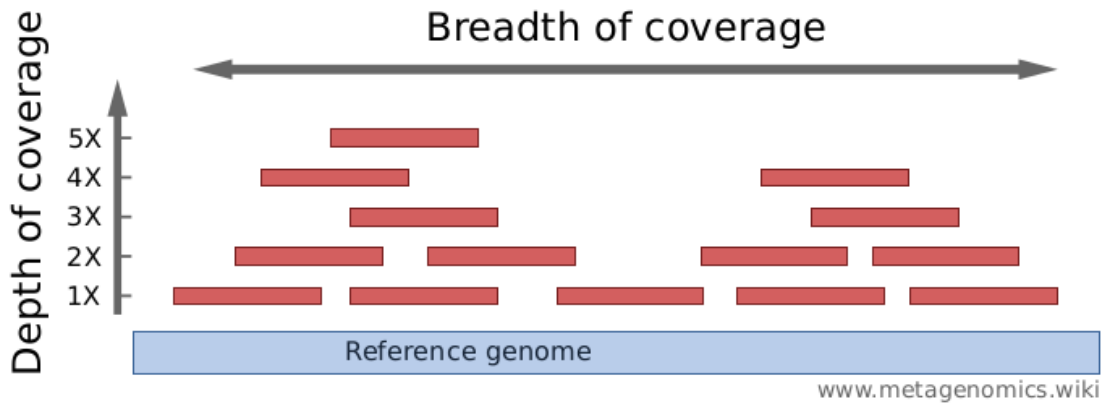
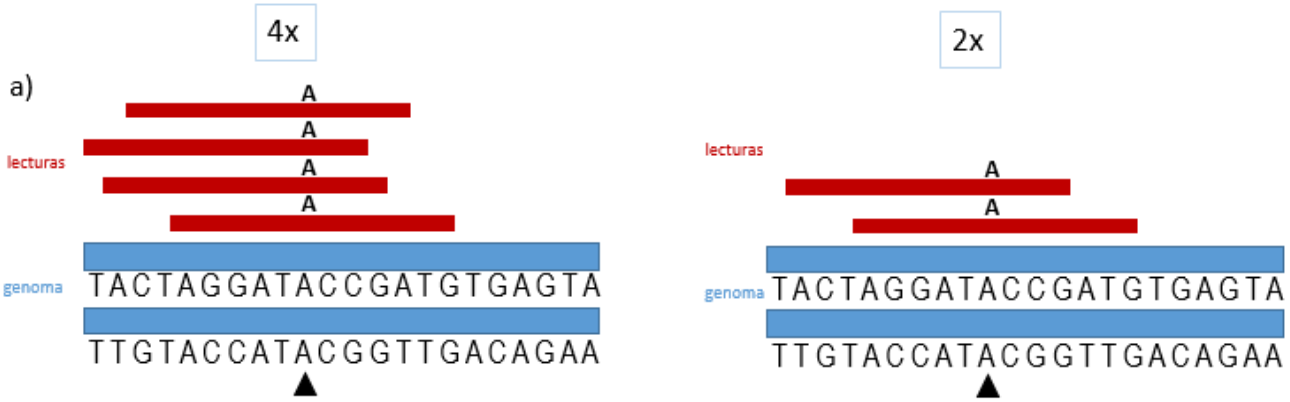


Figura 4. Secuenciación a una cobertura media de 3x.



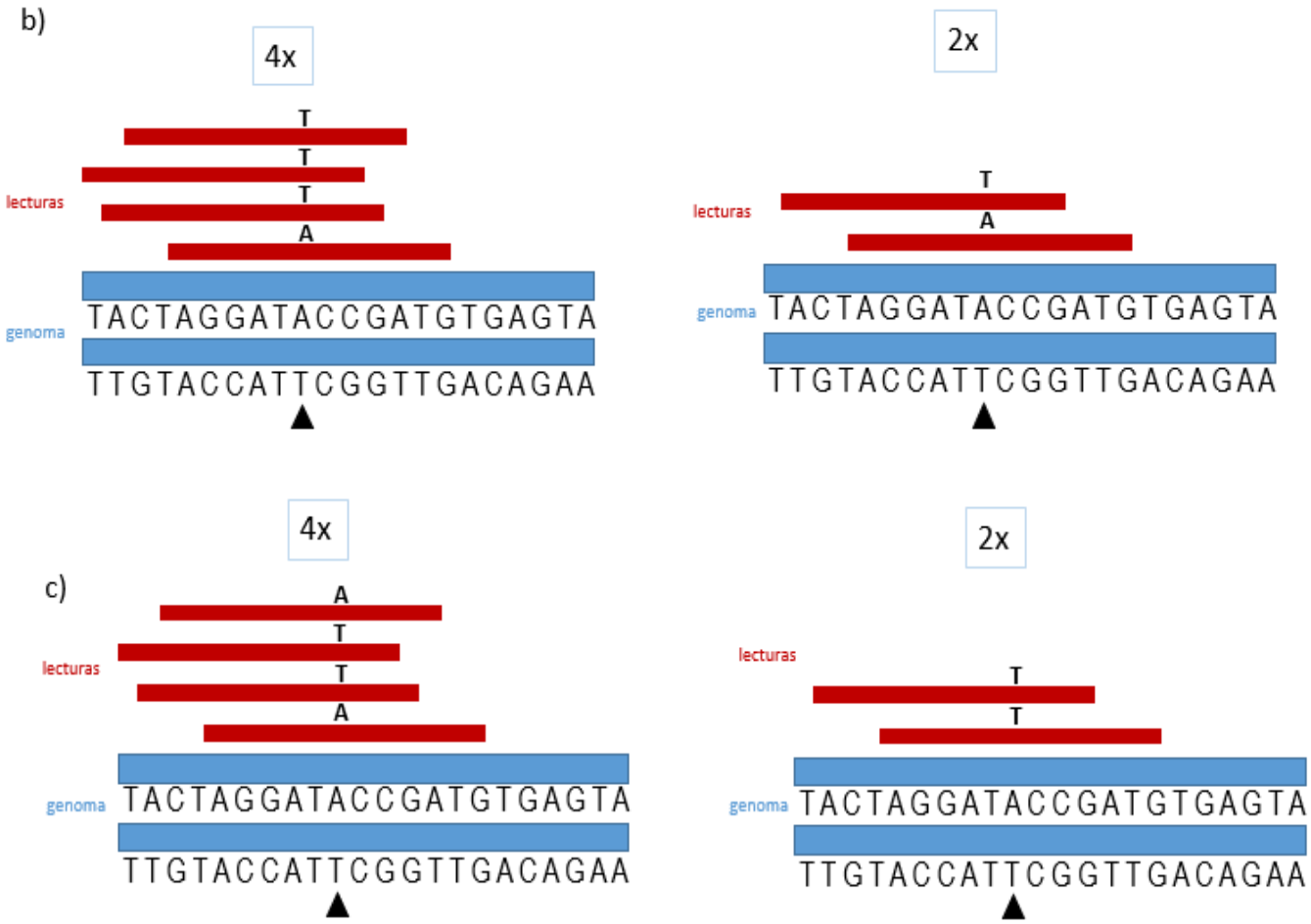


Figura 5. Ejemplo de la secuenciación como proceso de muestreo aleatorio de alelos, realizado con tres animales diferentes. En el lado izquierdo se muestra una cobertura de 4x y en el derecho de 2x. El animal a es homocigoto AA, por lo que tanto si muestreamos 4 lecturas como si muestreamos con 2 obtenemos siempre lecturas del alelo A. Los animales b y c son heterocigotos AT, por lo que realizando únicamente 2 lecturas podríamos observar 2 veces el alelo A, 2 veces T o 1 vez cada alelo. Por eso, a 2x en el animal b acertaríamos con el genotipo, en cambio, en el animal c cometeríamos un error, al no observar el alelo A y categorizarlo como un homocigoto. A medida que se dispone de más lecturas hay más probabilidades de observar los dos alelos y disminuye la incertidumbre del genotipo. Hay que recordar que además puede haber errores de secuenciación.

A veces es necesario realizar una secuenciación a coberturas muy elevadas; por ejemplo, en medicina humana es imprescindible para hacer diagnósticos individuales y fiables, ya que no se puede correr riesgos de equivocación. En estudios genómicos de animales (en este caso, cerdos) para la detección e identificación de genes asociados a caracteres productivos, una cobertura de 7x se puede considerar una cobertura suficiente para poder extraer buenos resultados a nivel poblacional, valores genéticos sobre la población (Ros-Freixedes et al., 2018). Se puede trabajar a coberturas menores, siempre y cuando se conozcan los errores y los sesgos que se producen en los genotipos obtenidos por medio de secuenciación a bajas coberturas (Roger et al., 2018). Una estrategia muy importante consiste en reducir la cobertura de secuenciación con el fin de reducir los costes de secuenciación individual (Hickey et al., 2013; Ros-Freixedes et

al., 2017). Aunque a medida que baja la cobertura de secuenciación se reduce el índice de detección de variantes a nivel individual, a nivel poblacional el índice de detección de variantes aumenta, al aumentar los individuos estudiados (Ros-Freixedes et al., 2020b). No obstante, aumenta la incertidumbre del genotipo (Ros-Freixedes et al., 2018), lo cual puede penalizar la potencia estadística de los estudios de asociación. Por lo tanto, es necesario evaluar tanto el índice de detección de variantes como la precisión de los genotipos obtenidos con esta tecnología a bajas coberturas.

Los estudios de asociación y la predicción genómica también se ven afectados por la cobertura, a coberturas superiores se obtienen resultados más fiables en ambos casos.

La secuenciación tiene como reto permitir la detección de variantes, que hasta ahora por medio de otros métodos no habían podido ser detectadas, aprovechando que estas nuevas técnicas tienen una mayor resolución (Ros-Freixedes et al., 2018) y de esta forma mejorar las predicciones genéticas y aumentar la precisión en los estudios de asociación.

1.3 Implementación práctica

Para que los datos de genotipado, sean de chips de genotipado o de secuenciación, se puedan usar de forma rutinaria en programas de mejora genética animal se deben adoptar medidas que reduzcan los costes de genotipado, como reducir la densidad o la cobertura y utilizar técnicas de imputación. La imputación consiste en genotipar para un número elevado de marcadores solo un subconjunto de individuos pertenecientes a una misma población y a partir de estos estimar los genotipos faltantes de los individuos del resto de la población, que se genotipan para un número mucho menor de marcadores. La imputación se basa en la similitud de los genomas de los individuos que pertenecen a una misma población, especialmente si están conectados mediante la genealogía.

Por ello, el diseño de programas de genotipado es complejo, ya que deben estimarse diferentes alternativas y escoger la que mejor se adapte al objetivo del estudio. Se debe tener en cuenta, entre otros factores, cual es la tecnología más adecuada para el genotipado, el número de individuos a genotipar, la estructura y genealogía de la población, la estrategia para seleccionar qué individuos genotipar, si se planea usar técnicas de imputación y cuales, y el presupuesto del que se dispone.

En este trabajo se intentará dar respuesta, utilizando diversos supuestos, a algunos de los principales aspectos del diseño de un programa de genotipado, como la elección de la tecnología de genotipado (chips de genotipado, secuenciación o una combinación de ambos), y la densidad, en el caso de los chips de genotipado, o la cobertura, en el caso de la secuenciación, en situaciones donde el presupuesto es un factor limitante.

2. OBJETIVOS

El objetivo general del trabajo ha sido investigar las oportunidades que ofrecen los chips de genotipado y la secuenciación completa del genoma como estrategias para la identificación de variantes asociadas a caracteres de interés en ganadería. Para ello, y utilizando como ejemplo un conjunto de marcadores de un chip asociados con el pH de la carne, se plantearon los siguientes objetivos específicos:

1. Detectar marcadores relacionados con el pH en cerdos Duroc por un estudio de asociación.

2. Validar la técnica de secuenciación en relación con los chips de genotipado:

   2.1 Determinar la capacidad de detección de los marcadores por secuenciación a diferentes coberturas.

   2.2. Determinar la concordancia genotípica entre los genotipos obtenidos por chips comerciales y por secuenciación.

   2.3 Validar la aplicación de secuenciación a coberturas bajas (máximo: 7x) como una tecnología viable para la identificación de variantes a nivel individual o poblacional.

3. Estudiar diferentes alternativas de genotipado en condiciones comerciales:

   3.1. Evaluar la proporción óptima de la población a genotipar con un chip y valorar el efecto de la densidad del chip en la precisión de la predicción.

   3.2. Evaluar la proporción óptima de la población a genotipar con un chip o por secuenciación bajo distintos presupuestos de inversión en genotipado.

3. MATERIALES Y MÉTODOS

3.1 Animales y datos

Se trabajó con datos de 395 cerdos Duroc de pura raza, seleccionada para la producción de carne de cerdo de calidad (Ros-Freixedes et al., 2012). Los animales muestreados pertenecían a 19 lotes productivos distintos. Aproximadamente a los 75 días después de su nacimiento, se trasladaron a las unidades de engorde (de 8 a 12 cerdos/corral), donde disponían de dietas ad libitum. Estos animales permanecieron allí hasta el sacrificio, siendo sacrificados en un matadero comercial. Después de 24 h del sacrificio, se midió el pH final en el músculo semimembranoso.

Los genotipos de estos animales se obtuvieron de dos maneras, por medio de chips de genotipado y por secuenciación, tal como se indica en los apartados 3.1.1 y 3.1.2, respectivamente. El ADN se extrajo de muestras de musculo semimembranoso usando el kit BioSprint DNA (Quiagen, Hilden, Germany). Los 395 cerdos fueron genotipados con chip y en 40 de ellos además se secuenció el genoma completo.

3.1.1 Chip de genotipado

Los datos de chip de genotipado fueron obtenidos a una densidad de 70.000 marcadores, usando los Chips GGP-Porcine HD BeadChip (GeneSeek, Lincoln, NE). A los marcadores se les sometió a un control de calidad, donde se eliminaron todos los marcadores que poseían una frecuencia alélica menor a 0.05. Debido al control de calidad y a que no se identificó a qué cromosoma pertenecían algunos marcadores, el número final de marcadores utilizado fue de 58.511 marcadores.

3.1.2 Secuenciación

Las muestras de los 40 animales se secuenciaron a una cobertura media de 6.8x (desviación estándar=1.2x; mín=4.4x; máx=12.2x). Se prepararon librerías de 300 pb que fueron secuenciadas con la técnica de extremos pareados (paired-end) usando NovaSeq 6000 (Illumina, San Diego, CA) en el CNAG (Barcelona). Las lecturas se proveen en un archivo con formato FastQ. Estas lecturas recibidas fueron alineadas al genoma de referencia Sscrofa11.1 (GenBank: GCA_000003025.6), utilizando el algoritmo BWA-MEM, por el cual se obtiene un archivo BAM. Este archivo BAM contiene las lecturas ordenadas según su posición relativa al genoma de referencia y parámetros de calidad según la certidumbre de este alineamiento (Mordoh, 2019). Las lecturas duplicadas se marcaron para ser excluidas usando la herramienta MarkDuplicates, de Picard (http://broadinstitute.github.io/picard/).

3.1.2.1 Variantes descubiertas

Una vez obtenido el archivo BAM, se trabajó con diversas herramientas de GATK. GATK es un kit de herramientas de análisis del genoma. Las variantes fueron detectadas mediante GATK HaplotypeCaller 3.8.0 (DePristo et al., 2011; Poplin et al., 2018), obteniendo un archivo gVCF. Este archivo es individual para cada uno de los animales, consta de todas las regiones estudiadas, tanto las que han mostrado variaciones como las que no. Posteriormente se utiliza la herramienta de GenotypeGVCFs, mediante la cual se obtuvieron los archivos VCF (variant call format). En un único archivo VCF se almacenan todas las variaciones presentes en los 40 animales a lo largo de la secuencia genética. Únicamente se estudiaron los SNP´s, de tal manera que no se tuvieron en cuenta las inserciones ni delecciones en el genoma. El número de lecturas que contenían el alelo de referencia (nRef) y el alelo alternativo (nAlt) se extrajeron usando una función de apilamiento (pile-up), siguiendo las recomendaciones de Ros-Freixedes et al. (2018) para evitar sesgos debidos a la baja cobertura (Figura 6).

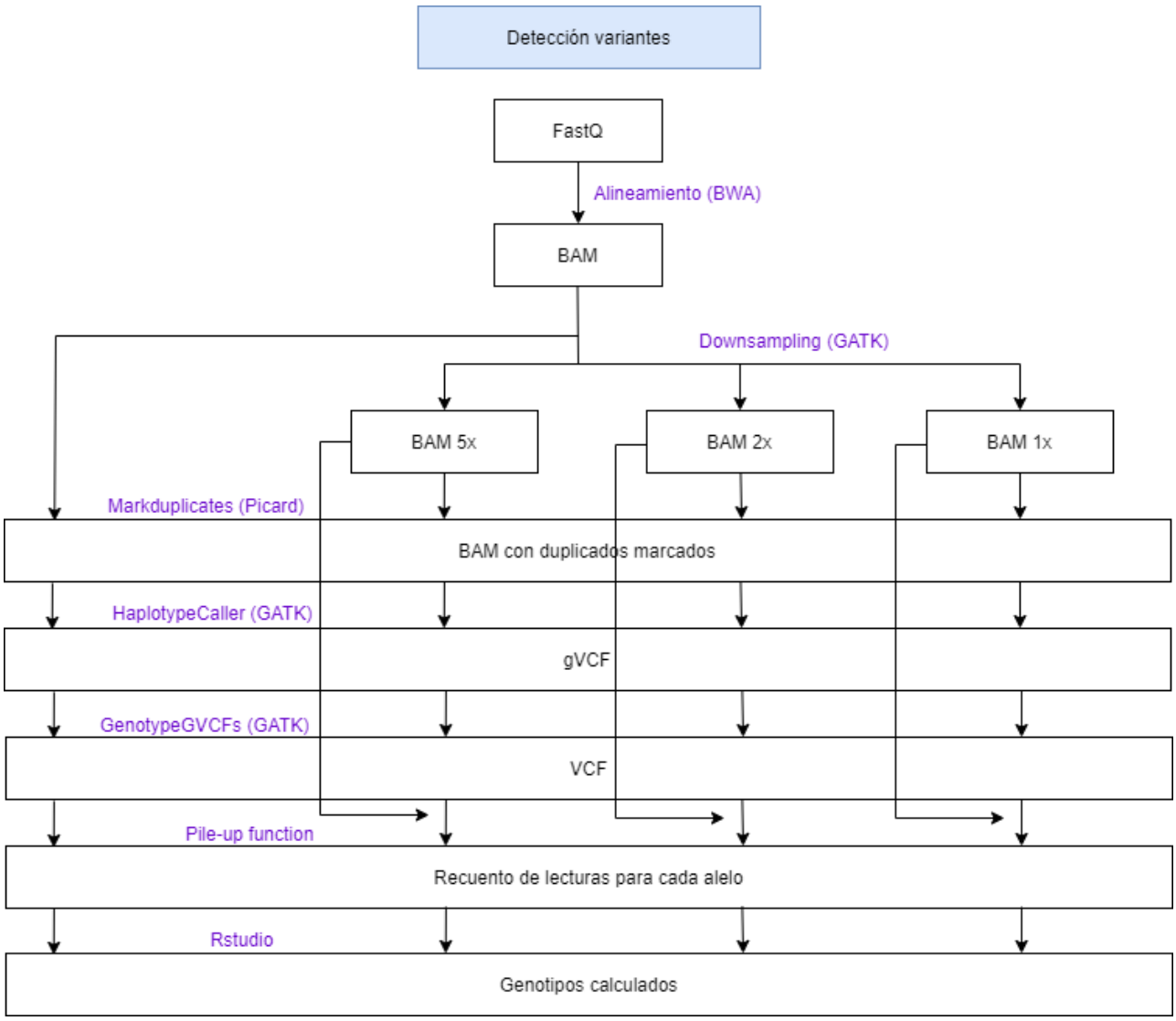


Figura 6. Esquema de los procesos durante la obtención de los genotipos. En color violeta se encuentran indicadas las herramientas bioinformáticas utilizadas en el proceso.

3.1.2.2 Genotipado

Se determinó el genotipo más probable para cada variante (SNP) aplicando las fórmulas que se muestran a continuación sobre el recuento de lecturas con el alelo de referencia (nRef) y con el alelo alternativo (nAlt). La nomenclatura utilizada para indicar al homocigoto de referencia fue el número 0; para el heterocigoto se utilizó un 1 y para el homocigoto alternativo un 2. Se siguió la metodología explicada en Ros-Freixedes et al., (2018).

El genotipo de cada animal para cada variante se calculó a partir de las probabilidades genotípicas (*p*) calculadas a partir del número de lecturas que contenían cada uno de los dos alelos posibles:

$$p\,(0) = (1-e)^{nRef} \times e^{nAlt}$$

$$p\,(1) = 0.5^{nRef} \times 0.5^{nAlt}$$

$$p\,(2) = e^{nRef} \times (1-e)^{nAlt}$$

Donde *e* es la tasa de error, que en este estudio se asumió de 0.001. Las probabilidades fueron ajustadas para que entre las tres sumaran 1. En todos los casos se escogió el genotipo con mayor probabilidad.

3.1.2.3 Reducción cobertura

Se analizó la capacidad de concordancia de genotipado comparando los genotipos obtenidos a diferentes coberturas. Para ello se llevó a cabo la reducción de la cobertura de forma aleatoria. Una vez que las lecturas fueron alineadas entre ellas y mapeadas con el genoma de referencia, por medio de las herramientas bioinformáticas descritas en 3.1.2.1, se redujo la cobertura para los datos de secuenciación de 7x a 5x, 2x y 1x. Para la reducción de los archivos se partió de los datos del archivo BAM (7x). Por medio de la herramienta GATK, con la función de Downsample se eliminan un porcentaje de las lecturas, escogidas éstas aleatoriamente. El porcentaje de las lecturas a eliminar se calculó en base a la cobertura inicial (alrededor de 7x) y la cobertura final deseada (1x, 2x, 5x) (Tabla 2).

Una vez se obtuvieron los archivos BAM a las coberturas requeridas, se identificaron los SNPs detectados en cada cobertura siguiendo el mismo proceso que se ha descrito en Ros-Freixedes et al., (2018).

Tabla 2: Coberturas obtenidas tras la eliminación de un porcentaje de lecturas aleatorias .

Cobertura esperada	Cobertura obtenida	Desviación estándar	Máximo	Mínimo
7x	6.8x	1.298	12.547	4.375
5x	5x	0.105	5.078	4.375
2x	2 x	0.008	2.042	2.008
1x	1x	0.004	1.024	1.007

Se observa que la única cobertura que es ligeramente distinta a la esperada es la de 7x, ya que, en las otras coberturas se eliminaron las lecturas de forma aleatoria por medio de GATK, indicando la cobertura que se requería al final de proceso. Por ello, la desviación estándar es pequeña en todos estos casos.

3.2 Estudio de asociación

Para demostrar la capacidad de identificación de variantes asociadas con un carácter, la información extraída por medio de los chips se analizó mediante un estudio de asociación (GWAS) sobre los 395 animales utilizando como ejemplo el carácter pH de la carne. Antes de realizar el GWAS el pH se corrigió por el efecto ambiental del lote productivo al que pertenecían los animales.

El objetivo de este estudio fue encontrar marcadores ligados al pH de la carne a 24 horas. Por medio del programa Plink (Chang et al., 2015; Purcell et al., 2007), se realizó un análisis GWAS, identificando y escogiendo los SNP más significativos asociados con el pH. Para ello, se calculó el p-valor asociado a cada uno de los marcadores y con ellos se construyó un gráfico Manhattan plot, que permite observar los cromosomas y las regiones donde se localizan los marcadores más ligados al fenotipo del pH.

Se extrajeron los marcadores con una asociación al pH más significativa (menor p-valor) con la ayuda de Rstudio. Se utilizó la herramienta Biomart de Ensembl (Yates et al., 2020) para identificar los genes localizados en las regiones más asociadas con el pH y la herramienta Enrichr (Chen et al., 2013; Kuleshov et al., 2016) para determinar su función biológica.

3.3 Validación de la secuenciación a coberturas bajas y moderadas

Por otro lado, para calcular la validación del genotipado mediante secuenciación y la concordancia genotípica (concordancia entre los genotipos obtenidos por medio de chips de genotipado con los conseguidos por la secuenciación), se trabajó con los 40 animales que disponían datos de secuenciación.

Realizamos un GWAS con los datos de chip sobre estos 40 individuos, los marcadores obtenidos tras realizar el mismo control de calidad que el descrito en 3.1.1, fueron 45.093 SNPs. Los marcadores que se compararon fueron los que presentaron un valor del –log del p-valor superior a 4 en el GWAS descrito en 3.2. y que segregaban en los 40 animales secuenciados. Muchos de estos marcadores pertenecían a una misma región del genoma, por lo que se escogió solo un marcador de estas regiones. El número final de marcadores estudiado fue 16.

Se analizó la capacidad de detección de variantes mediante secuenciación, al comparar el número de variantes detectadas a 5x, 2x y 1x con las obtenidas a 7x, asimismo respecto a las variantes presentes en el chip y sobre los 16 marcadores seleccionados. También la concordancia entre los genotipos y los alelos obtenidos por chips y secuenciación a diferentes coberturas (7x, 5x, 2x y 1x). Se asumió que los genotipos de referencia eran los procedentes de los chips, a pesar de que también pueden contener

errores. La concordancia genotípica se calculó como el porcentaje de coincidencias entre los genotipos obtenidos por el chip con los genotipos obtenidos para cada cobertura. La concordancia alélica se extrajo de la misma forma, pero contando los alelos coincidentes (es decir, la concordancia alélica entre un homocigoto y un heterocigoto es 0.5).

3.4 Implementación práctica

3.4.1 Escenarios

Se realizó un estudio económico de diferentes escenarios con el fin de encontrar aquella estrategia de genotipado que combinada con imputación proporcione mayor precisión de genotipado de una población entera al menor coste posible. Los escenarios se configuraron variando los siguientes factores: 1) densidad del chip; 2) proporción de la población genotipada con chip; 3) estrategia para asignar la densidad del chip de genotipado a cada animal según su progenie; 4) cobertura de secuenciación; 5) porcentaje del presupuesto total destinado a chips o secuenciación; y 6) presupuesto total disponible para genotipado.

En todos los escenarios se simularon los genotipos de una población de 30.000 animales estructurados en torno a un pedigrí real extraído de un núcleo de selección cerrado con varias generaciones continuas y solapadas. Se simuló un cromosoma de 100 cM de tamaño con 24.000 SNPs. Para cada individuo se simularon genotipos para estos 24.000 SNPs. Asumimos que los cerdos se podían genotipar con dos tipos de chips, uno de alta densidad 55k (HD) y otro de baja densidad 7k (LD), de manera que en los SNPs incluidos en los chips HD y LD se seleccionaron como un subconjunto de 3,000 y 300 marcadores elegidos al azar y anidados, los cuales equivaldrían a los SNPs pertenecientes a un cromosoma, para conseguir de esta forma que los datos simulados reflejaran una estructura de datos reales.

Se realizaron dos estudios o tests analizando diferentes parámetros, estrategias de genotipado e inversión en las tecnologías de genotipado, creando 16 escenarios y 54 respectivamente. Estos estudios económicos se realizaron con el fin de obtener datos a nivel poblacional, no individual.

3.4.2 Costes de las tecnologías utilizadas

Los precios de los métodos de secuenciación utilizados son un reflejo de los costes de mercado actuales. En la secuenciación se asumió un coste fijo de 40 € para crear las librerías, más un coste lineal de 32 € por cada unidad de cobertura (Tabla 3).

Tabla 3: Costes de chips de genotipado y de secuenciación.

Método	Densidad	Cobertura	Coste (€)
Chip HD	55K	-	40
Chip LD	7K	-	15
Secuenciación	-	7x	300
Secuenciación	-	5x	200
Secuenciación	-	2x	104
Secuenciación	-	1x	72

3.4.3 Primer test: Estrategias de genotipado

En el primer test se evaluaron 16 escenarios, para comprobar las mejores estrategias de genotipado; el porcentaje de la población genotipada por chip (100% o 50%), el tipo de chip utilizado (HD o LD) en función del número de descendientes, y la cobertura de secuenciación (7x, 5x ,2x o 1x). Se utilizó un presupuesto único de 600.000€ en todos los escenarios. En cada uno de los escenarios se combina una estrategia de genotipado mediante chips y una estrategia de secuenciación, tal como se indica en la Tabla 4.

Tabla 4. Los distintos escenarios, donde se muestra el número de animales genotipados con chips a alta densidad (HD), baja densidad (LD) y secuenciados.

	HD100%	HD50%	LD100%	LD50%
1x	117 HD; 29883 LD; 2042 secuenciados	117 HD; 14883 LD; 5167 secuenciados	30000 LD; 2083 secuenciados	15000 LD; 5208 secuenciados
2x	117 HD; 29883 LD; 1414 secuenciados	117 HD; 14883 LD; 3577 secuenciados	30000 LD; 1442 secuenciados	15000 LD; 3605 secuenciados
5x	117 HD; 29883 LD; 735 secuenciados	117 HD; 14883 LD; 1860 secuenciados	30000 LD; 750 secuenciados	15000 LD; 1875 secuenciados
7x	117 HD; 29883 LD; 490 secuenciados	117 HD; 14883 LD; 1240 secuenciados	30000 LD; 500 secuenciados	15000 LD; 1250 secuenciados

Se testaron cuatro estrategias de genotipado: HD100%, HD50%, LD100% y LD50%. En HD100% se genotiparon los 117 animales con más progenie (hijos de más respecto a la progenie media) mediante un chip HD mientras que el resto (100%) lo fueron con un chip LD. En HD50% se genotiparon los 117 animales con más progenie mediante un chip HD y el 50% del resto de la población (escogida aleatoriamente) con LD. En LD100% se genotipo el 100% de los individuos con un chip LD mientras que en LD50% únicamente el 50% de la población (escogida aleatoriamente) se genotipó con un chip LD.

Una vez cubiertos los gastos del genotipado por chip, el capital restante se usó para secuenciación. El número de animales a secuenciar se determinó a partir del presupuesto disponible y la cobertura de secuenciación. Los animales a secuenciar se escogieron de manera aleatoria. Finalmente se estudiaron las concordancias obtenidas para cada cobertura 7x, 5x, 2x, 1x.

3.4.4 Segundo test: Inversión en las tecnologías de genotipado

Una vez evaluado el porcentaje de la población, la mejor cobertura de secuenciación y el tipo de chip óptimo para los animales con alta progenie, se llevó a cabo un segundo estudio. En este análisis se trabajó variando los presupuestos. Se crearon 54 escenarios. Las variables que se examinaron fueron:

1) El porcentaje de capital destinado a cada tecnología (chip, secuenciación). El rango invertido en estas tecnologías fue del 10%, 20%, 30% …. 90%.
2) Presupuesto disponible. Los presupuestos trabajados fueron 150.000€, 300.000€, 450.000€, 600.000€, 750.000€, 900.000€.

3.4.5 Datos simulados

Los datos fueron simulados por medio de AlphaSimR (Faux et al., 2016) creándose un total de 24.000 SNPs por cromosoma. Se creó un subconjunto aleatorio de 5.000 SNPs utilizado como una representación de los marcadores de secuenciación, para reducir el coste computacional del análisis.

Las lecturas obtenidas en la secuenciación se simularon una vez se habían seleccionado de forma aleatoria los individuos a secuenciar y la cobertura requerida. Se simuló el recuento de lecturas que contienen cada alelo en cada individuo, para reproducir el muestreo aleatorio de alelos como ocurre en las plataformas de secuenciación.

3.4.6 Imputación

Se asumió que para los animales genotipados por el chip solo se conocían los genotipos en los SNPs incluidos en el chip correspondiente pero no el resto, y en los animales secuenciados se asumió que solo se conocía el recuento de lecturas para cada alelo obtenidas mediante un muestreo aleatorio para cada cobertura. Se usó esta información para imputar los genotipos en las posiciones restantes donde no se conocía el genotipo. La imputación se realizó utilizando la herramienta AlphaPeel (Whalen et al., 2018) con los parámetros por defecto. Este programa realiza dos pasos. En un primer paso se usaron los genotipos obtenidos mediante chips de genotipado para determinar la probabilidad de que el genotipo de cada uno de los 30.000 individuos proviniera de cada uno de sus haplotipos parentales, usando un algoritmo iterativo basado en la genealogía. En un segundo paso se usaron estas estimaciones sobre los recuentos de las lecturas en cada alelo para determinar el genotipo más probable para cada animal teniendo en cuenta la información de toda la genealogía.

La precisión de la imputación se midió como la concordancia entre los genotipos verdaderos (simulados) y los obtenidos por la imputación.

4. RESULTADOS

4.1 GWAS

Se realizó un estudio de asociación sobre los 395 animales para el carácter del pH. Se aplicó la corrección de Bonferroni para múltiples tests. Los resultados se muestran en un 'Manhattan plot' (Figura 7), que nos indica el p-valor de los marcadores asociados con el pH y su posición en el genoma.

El umbral de significación -log (p-valor) corregido por Bonferroni fue de 6. En el estudio no se detectó ninguna asociación con p-valor superior a este umbral, por lo que tras interpretar visualmente el Manhattan plot se decidió estudiar como "SNPs sugestivos" aquellos que poseían un –log (p-valor) igual o superior a 4.

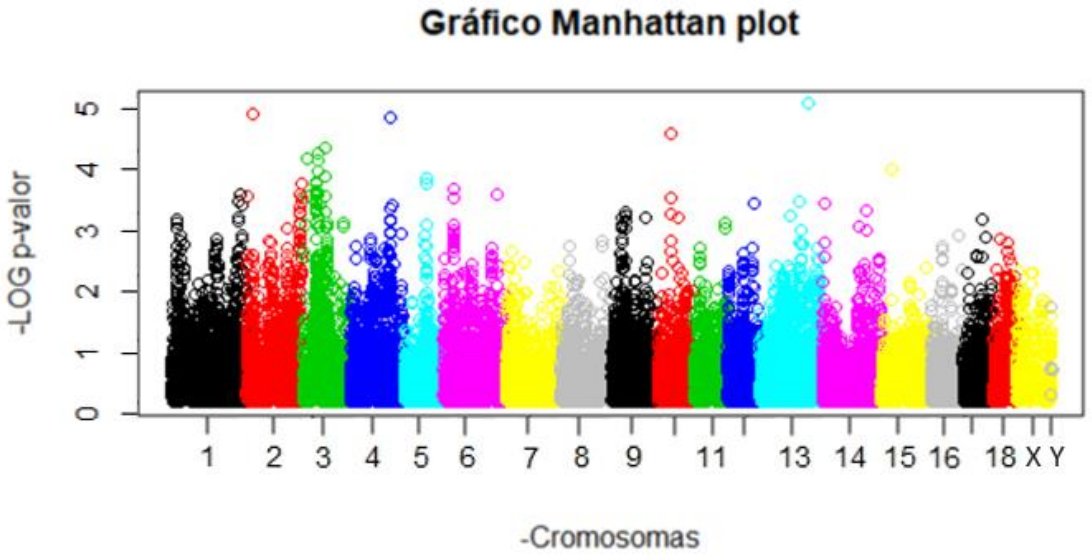


Figura 7. Manhattan plot sobre el pH de la carne. Los diferentes colores indican los diferentes cromosomas presentes en el cerdo. En el eje de las Y se localizan los diferentes valores del –log p-valor y en el eje de las X indica la posición en pares de bases dentro de cada uno de los cromosomas

Una vez escogidos los SNPs con mayor asociación con el pH se tomó un rango de 1 Mb alrededor suyo para detectar los genes presentes en cada una de las regiones (Tabla 5).

Tabla 5: Genes presentes en las regiones ligadas al carácter del pH.

Cromosoma	Posición (pb)	-log (p-valor)	Genes
10	21325625:22325625	4,50	*ATP6V1G3, PTPRC*
13	176886304:177886304	5,00	*ROBO2*
2	11216316:12216316	4,82	*MS4A5, MS4A2, OOSP3, TCN1, CBLIF, MRPL16, STX3, OR10V1, PATL1, U6, OR4D10, MPEG1, DTX4*
4	104293040:105293040	4,76	*ATP1A1, MAB21L3, SLC22A15, NHLH2, CASQ2, VANGL1, NGF*

3	63164416:64164416	4,25	-
3	27499660:30212394	4,17	*U6, ABCC1, MRTFB;ERCC4, SHISA9, U6, LRRTM1*
3	6513004:7513004	4,07	*ZCAN25, CYP3A29, FOPNL, MYH11, NDE1 MARF1, BMERB1, PDXDC1, MPV17L*

Los genes presentados en la Tabla 5 son aquellos genes candidatos que potencialmente pueden relacionarse con el pH del músculo. Las funciones biológicas principales de estos genes son actividades metabólicas, de transporte transmembrana y de regulación. Al observar las funciones moleculares se detectó que había entre ellos el gen *ATP1A1*, que podía relacionarse directamente con el pH, ya que afectaba al transporte de potasio, a la actividad de intercambio potasio/sodio y a la unión de sodio (Figura 8). Otros genes candidato podrían ser el *ABCC1* y *SLC22A15*, pero parecían estar más relacionados con el transporte de aniones orgánicos que no directamente con el pH.

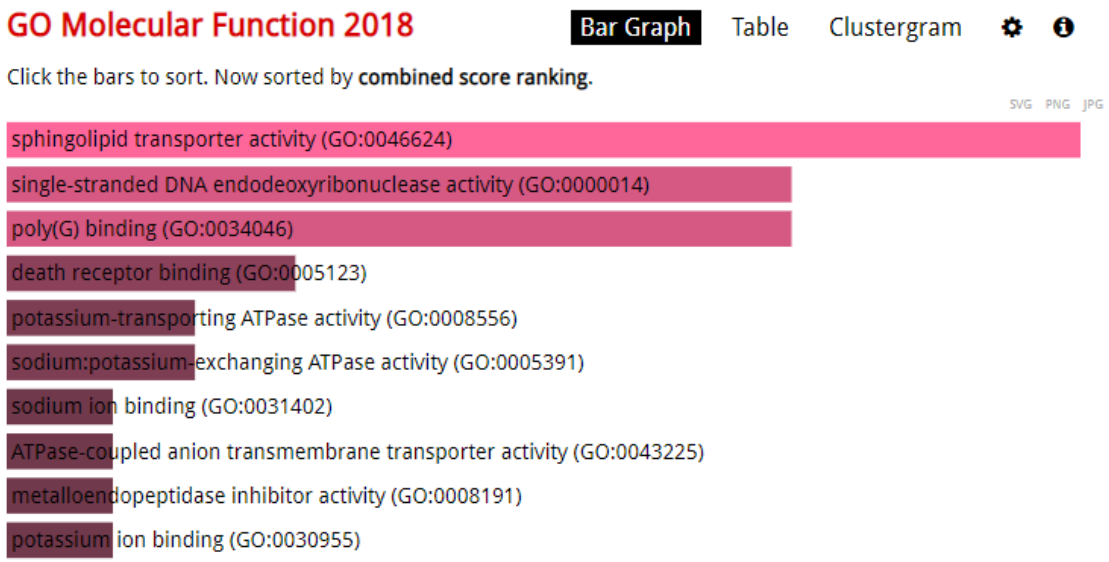


Figura 8. Gráfico de las funciones moleculares obtenidas de Enrichr sobre los genes adquiridos por medio de la herramienta Biomart de Ensembl.

4.2 Comparación genotipos

Como se disponía únicamente de 40 animales con datos de secuenciación a una cobertura de 7x, escogimos 16 marcadores que segregaran para estos individuos. Se realizaron estudios de concordancia entre los chips de genotipado y la secuenciación con estos marcadores.

Las variantes descubiertas en cada cobertura se muestran en la Tabla 6. Se observa una reducción de las variantes detectadas al disminuir la cobertura de secuenciación tanto en las variantes totales de secuenciación, como en las del chip y en los 16 marcadores seleccionados.

Tabla 6: Variantes identificadas mediante secuenciación para cada una de las coberturas. El porcentaje muestra el porcentaje de variantes detectadas respecto a las obtenidas en secuenciación a 7x, en los chips de genotipado y en los 16 marcadores escogidos.

	Nº variantes totales	Nº variantes chip (45,093 SNPS)	Nº variantes marcadores (16 marcadores)
7x	15,941.705 (100%)	42,954 (95.25%)	16 (100%)
5x	15,342.897 (96.24%)	42,922 (95.18%)	16 (100%)
2x	13,247.087 (83.10%)	41,857 (92.82%)	15 (93.75%)
1x	10,974.412 (68.84%)	37,644 (83.48%)	14 (87.5%)

Las concordancias genotípicas para cada cobertura se muestran en la Tabla 7, donde también se indica si el marcador fue detectado para cada una de las coberturas de secuenciación.

Tabla 7. Concordancia genotipos en las diferentes coberturas estudiadas. Con un guion se muestran aquellas posiciones en las que no fue posible detectar la variante.

	Marcador	Cromosoma y posición	Detectado	Concordancia genotipos (7x)	Concordancia genotipos (5x)	Concordancia de genotipos (2x)	Concordancia de genotipos (1x)
1	rs80886405	1_22728315	Sí	95,0%	92,5%	82,5%	60,0%
2	rs339716177	4_120993942	Sí	50,0%	47,5%	45,0%	30,0%
3	rs81388525	6_72030654	Sí	97,5%	95,0%	70,0%	52,5%
4	rs322695071	7_9581427	Sí	92,5%	90,0%	60,0%	27,5%
5	rs81406039	8_137284238	Sí	97,5%	92,5%	60,0%	57,5%
6	rs81345873	9_131683136	Sí	97,5%	90,0%	62,5%	35,0%
7	rs81266524	10_45722031	Sí	90,0%	87,5%	72,5%	32,5%
8	rs333143010	11_75533249	Sí	62,5%	65,0%	50,0%	27,5%
9	rs81430187	11_19677423	Sí	55,0%	52,5%	32,5%	20,0%
10	rs81327834	12_44051828	Sí	42,5%	42,5%	30,0%	35,0%
11	rs340756862	13_199441131	Sí	92,5%	95,0%	75,0%	47,5%
12	ALGA0123146	14_1243912	Sí	97,5%	95,0%	82,5%	52,5%
13	rs323909695	15_122205759	En 1x no,	75,0%	75,0%	62,5%	-
14	rs81258713	17_44751346	En 1x no, en 2x no	92,5%	90,0%	-	-
15	rs341071049	X_8152258	Sí	87,5%	77,5%	67,5%	35,0%
16	rs324254306	X_ 3311527	Sí	95,0%	92,5%	70,0%	42,5%

Para algunos marcadores se obtuvieron concordancias muy elevadas como en los marcadores 3, 5, 6 y 12. En cambio, para la misma cobertura otros marcadores como el 2, 8, 9 y 10 presentaron concordancias muy bajas.

Los marcadores 13 y 14 no fueron detectados con las coberturas más bajas. Por otro lado, los marcadores 15 y 16 se encontraban en el cromosoma X, por lo que la cobertura a la que fueron obtenidos estos genotipos era la mitad, ya que, los animales secuenciados eran machos.

Se realizó un estudio de concordancia alélica con el fin de comprobar si los genotipos erróneos se debían al cambio en un alelo o a los dos. Los resultados se muestran en la Tabla 8.

Tabla 8. Concordancia alelos en las diferentes coberturas. Con un guion se muestran aquellas posiciones en las que no fue posible detectar la variante.

	Marcador	Cromosoma y posición	Detectado	Concordancia alelos (7x)	Concordancia alelos (5x)	Concordancia alelos (2x)	Concordancia alelos (1x)
1	rs80886405	1_22728315	Sí	97,50%	95,00%	87,50%	62,50%
2	rs339716177	4_120993942	Sí	50,00%	48,75%	62,50%	38,75%
3	rs81388525	6_72030654	Sí	98,75%	97,50%	80,00%	61,25%
4	rs322695071	7_9581427	Sí	96,25%	95,00%	72,50%	37,50%
5	rs81406039	8_137284238	Sí	98,75%	96,25%	75,00%	63,75%
6	rs81345873	9_131683136	Sí	98,75%	95,00%	71,25%	45,00%
7	rs81266524	10_45722031	Sí	95,00%	93,75%	81,25%	41,25%
8	rs333143010	11_75533249	Sí	75,50%	78,75%	63,75%	42,50%
9	rs81430187	11_19677423	Sí	75,00%	72,50%	56,25%	32,50%
10	rs81327834	12_44051828	Sí	43,75%	45.00%	40,00%	48,75%
11	rs340756862	13_199441131	Sí	93,75%	95,00%	80,00%	52,50%
12	ALGA0123146	14_1243912	Sí	98,75%	97,50%	87,50%	55,00%
13	rs323909695	15_122205759	En 1x no,	76,25%	77,50%	70,00%	-
14	rs81258713	17_44751346	En 1x no, en 2x no	96,25%	95,00%	-	-
15	rs341071049	X_8152258	Sí	88,75%	78,75%	67,50%	35,00%
16	rs324254306	X_ 3311527	Sí	97,50%	93,75%	70,00%	42,50%

Con los resultados obtenidos se observó que a una cobertura de 5x los resultados eran similares a los que se consiguieron a cobertura de 7x. En cambio, al reducir la cobertura a 2x o 1x se produjo una importante pérdida de información genotípica.

4.3 Implementación económica

4.3.1 Primer test

Los resultados al variar los parámetros de este primer estudio se pueden observar en la Tabla 9:

Tabla 9: Precisión de la imputación en los diversos escenarios, donde se varió la secuenciación, porcentaje de la población genotipada por chip, y el tipo de chip utilizado dependiendo de la progenie de cada animal. HD y LD hace referencia al chip de alta y baja densidad, el porcentaje que acompaña al chip indica la población genotipada y la cobertura de secuenciación se indica por 1x, 2x, 5x y 7x.

Secuenciación	Chip			
	HD100%	HD50%	LD100%	LD50%
1x	93.86%	87.92%	94.20%	88.57%
2x	94.42%	88.80%	95.01%	88.93%
5x	93,90%	88.52%	94.63%	88.68%
7x	92.90%	87.98%	93.66%	88.21%

Los escenarios más favorables eran aquellos en los que se genotipaban a baja densidad toda la población (LD100%), ya que, de esta manera se recopilaba una mayor cantidad de información. Al variar la cobertura se afectaba la precisión en cada escenario, mostrándose su máximo en 2x. Al genotipar los animales que poseían una mayor progenie a HD en vez de LD, no se obtuvo mejor precisión.

4.3.2 Segundo test

La cobertura a la que se realizó este segundo test se fijó a 2x, ya que, por el primer test se comprobó que era la cobertura que proporcionaba una mejor precisión. Por este mismo motivo únicamente se utilizaron chips a baja densidad (LD).

Al cambiar los presupuestos se observó que el porcentaje óptimo de capital invertido en cada una de las tecnologías variaba (Tabla 10). En los presupuestos de 150.000€ y 300.000 no se detectó ninguna tendencia clara, a partir del presupuesto de 450.000 se observó que, al aumentar el presupuesto, la parte disponible para secuenciación aumentaba paulatinamente y la precisión de la imputación aumentaba. En la Tabla 10 se indica con un guion aquellos resultados que se mantienen, el presupuesto disponible supera el coste de genotipar toda la población con chips. En estos casos, los 30.000 animales poseen datos de chip, por lo que no tiene sentido seguir invirtiendo en chips y los porcentajes destinados a cada tecnología son los mismos.

Tabla 10: Resultados test dos. Concordancias genotípicas obtenidas al variar el presupuesto y el porcentaje de capital destinado a cada tecnología. Con un guion se muestran los resultados que se mantienen, debido a que los 30.000 individuos ya poseían datos de chip.

Presupuesto disponible	Porcentaje de presupuesto destinado a chips de genotipado								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
150.000 €	77,83%	78,35%	78,54%	78,80%	78,84%	79,42%	80,32%	80,89%	81,46%
300.000 €	80,14%	81,08%	82,25%	82,79%	83,71%	84,48%	85,05%	84,72%	83,23%
450.000 €	81,25%	83,16%	84,84%	86,25%	87,67%	88,95%	89,99%	90,39%	88,94%
600.000 €	83,07%	85,08%	87,26%	89,37%	91,10%	92,89%	94,34%	95,01%	-
750.000 €	84,30%	86,91%	89,45%	92,11%	94,43%	96,68%	96,68%	-	-
900.000 €	85,05%	88,63%	91,49%	94,47%	97,38%	96,48%	-	-	-

5. DISCUSIÓN

5.1 GWAS

Mediante el GWAS identificamos el gen candidato *ATP1A1* con funciones que podrían estar relacionadas con el pH (Tabla 5).

El gen *ATP1A1* se encuentra en el cromosoma 4 y está relacionado con el transporte de cationes como Na^+ / K^+. La proteína codificada por este gen pertenece a la familia de las ATPasas. Esta proteína se encarga de mantener los gradientes electroquímicos a través de la membrana por los iones Na^+ y K^+. Este gen junto a otros codifican para una enzima formada por dos subunidades: una subunidad catalítica α y otra subunidad de glucoproteina β, más pequeña. El *ATP1A1* es uno de los genes que afectan a la subunidad catalítica (Sahoo et al., 2016). La subunidad α juega un papel indispensable para la homeostasis de iones celulares (Schlingmann et al., 2018). Además, esta subunidad catalítica de la enzima predomina en los glóbulos rojos (estos son importantes amortiguadores ácido-base que mantienen el pH) y en el tejido nervioso (Sahoo et al., 2016).

Algunas mutaciones de *ATP1A1* producen una fuga de protones o sodio por vía intracelular. En humanos, algunas mutaciones como *ATP1A1* ^{L104R} o *ATP1A1* ^{V332G} mostraron una despolarización patológica de la membrana. En las mediciones del pH intracelular estas mutaciones causaron una acidificación intracelular, que provocó una capacidad reducida de las células para mantener constante el pH intracelular. Se ha descrito que la mitad de los pacientes con aldosteronismo primario (que es la forma más común de hipertensión secundaria) tenían esta enfermedad debido a una mutación en este gen. Algunas de estas mutaciones se encontraban en la subunidad α de la ATPasa de sodio y potasio y aumentaban la producción de aldosterona (Gomez-Sanchez et al., 2015). La aldosterona es una hormona esteroide que actúa en la conservación del sodio, tanto secretando potasio como incrementando la presión sanguínea. La perturbación de la homeostasis del pH posiblemente ayude a la producción autónoma de aldosterona (Stindl et al., 2015).

En diversas especies de ganado se han realizado análisis bioinformáticos comparativos de las proteínas *ATP1A1* y sus variantes genéticas. Se ha observado que estas variantes genéticas del gen *ATP1A1* se relacionaban con la hipertensión (Sahoo et al., 2016). Otro estudio en roedores indica que algunas de las mutaciones en este gen produce flujos de iones anormales, mientras que análisis realizados sobre el pH intracelular mostraron una permeabilidad anormal de H^+ y cambios significativos en el pH intracelular producidos por el genotipo mutante (Schlingmann et al., 2018).

Al comparar los resultados de nuestro GWAS sobre el pH muscular con los de Davoli et al (2019), también en Duroc, se comprobó que los genes detectados eran distintos. Esta diferencia puede deberse, por ejemplo, al distinto origen del material genético de las dos líneas o a diferencias en la selección realizada en ambas. De los cuatro marcadores más asociados en el estudio de Davoli et al., (2019), en nuestro chip de genotipado solo disponíamos del CASI0005117. Este no salió significativo en nuestra población. La única similitud entre ambos trabajos fue que en los dos destacaron regiones del

cromosoma 3; sin embargo, nosotros no observamos ningún marcador significativo en la región que ellos encontraron de interés (16963061:17963061).

Los estudios de asociación poseen limitaciones cuando una población se encuentra muy seleccionada, ya que, es posible que la alteración o los marcadores estén fijados en todos los individuos de esa población y no seamos capaces de detectarla por el GWAS. Otra de las limitaciones es que podemos perder mutaciones esporádicas (con una frecuencia muy pequeña) al considerar estas alteraciones como un error de computación y eliminarlos al realizar un control de calidad. En este trabajo todas las alteraciones que aparecieran en una frecuencia menor al 0.05 fueron eliminadas.

5.2 Comparación genotipos

Al realizar la comparación de los genotipos, se tomaron como referencia los obtenidos por los chips de genotipado y se contrastaron con los de secuenciación. Se consiguieron precisiones muy diversas según el marcador. Al observar que una parte de los marcadores poseían concordancias bajas, como los marcadores 2, 8, 9 y 10 (Tabla 7), se examinaron las lecturas obtenidas para cada alelo en las coberturas realizadas para estas posiciones y no encontramos ninguna evidencia de sesgo de alineamiento. Por este motivo no podemos descartar que la baja concordancia se deba a errores en los datos de secuenciación o errores en el chip.

Hubo genotipos que no se pudieron predecir, al no poseer lecturas suficientes. Estos valores sin genotipo fueron valorados como resultado erróneo. Por lo que las discordancias entre genotipos o alelos podía deberse a que no hubiera dato (siendo estos casos un 0,6%, 1,9%, 14,8% y 40,1% para 7x, 5x, 2x y 1x) o a que no hubiera concordancia genotípica o alélica. Por medio de la Tabla 7 se pudo comprobar que muchos de los genotipos que no concordaban se debían al cambio de un alelo, pasando de un homocigoto a un heterocigoto o viceversa. Había pocos genotipos en los que se presentaban homocigotos opuestos. Los errores entre homocigoto y heterocigoto tienen un menor impacto sobre las asociaciones marcador-carácter que los errores entre homocigotos opuestos, que son mucho más desfavorables.

Se detectaron con éxito 14 de los 16 marcadores en todas las coberturas. No se detectaron los marcadores 13 y 14 a bajas coberturas. Esto se debió a que no aparecía ningún tipo de variación en las lecturas para esas coberturas. Al ser coberturas bajas, los animales secuenciados no disponían de genotipo para esos marcadores, o bien, si lo tenían, era siempre el mismo.

La concordancia a una cobertura de 7x para la mayoría de los 16 marcadores fue alta, de mediana 92,50% y media 82,50%. Únicamente en unos pocos marcadores se obtuvieron concordancias bajas. Para 5x se consiguió una mediana de 90,00% y concordancia media de 80,00%. Se observa que las concordancias conseguidas en ambas coberturas no fueron muy diferentes. En cambio, a coberturas menores en 2x la mediana fue 62,50% y la concordancia media de 58,00%, mientras que en 1x se obtuvo un valor de 35,00% en ambos casos, provocando una gran cantidad de errores en la predicción de los genotipos. Por medio de estos valores se determinó que para realizar estudios a nivel individual se deberían realizar a coberturas de 7x o 5x, ya que, si no se cometerían

muchos errores. Una secuenciación a 5x proporciono un buen equilibrio entre las variantes descubiertas y la precisión. En un artículo de Jiang et al., (2019) se estudiaron doce coberturas distintas de secuenciación, con un rango de 20x a 1x, y determinaron que una cobertura inferior a 4x producía un aumento de genotipos falsos (especialmente a coberturas < 2x). También indican que con una cobertura de 4.35x se cubría más de 95% del genoma de un cerdo.

En estudios a nivel poblacional se puede disminuir la cobertura hasta 2x, siempre y cuando se disponga de información sobre un gran número de animales relacionados, se conozca su genealogía y se usen técnicas de imputación para usar toda la información disponible en la población de forma conjunta (Ros-Freixedes et al., 2020a,b). A nivel individual o cuando la imputación no es posible, a coberturas iguales o inferiores a 2x se pierde demasiada información. La imputación es una herramienta muy potente para equilibrar el coste y la eficiencia de la secuenciación y permite lograr buenas precisiones genotípicas (Jiang et al., 2019). La precisión de la imputación depende mucho de la profundidad y lo entrelazados que estén los individuos de una población. En el trabajo de Ros-Freixedes et al., 2020b se puede observar que se consiguieron precisiones altas secuenciando el 2% de la población a una cobertura a 2x. Con la ayuda de algoritmos de imputación específicos para secuenciación a baja cobertura, Jiang et al. (2019) consiguieron una precisión en los genotipos de los ratones mayor al 90% a coberturas < 1x.

5.3 Implementación práctica

5.3.1 Primer test

La mejor cobertura de secuenciación a nivel poblacional cuando se aplicaron técnicas de imputación se obtuvo en torno a 2x. Esta cobertura permite aumentar el número de animales secuenciados, debido a que el coste es menor que a coberturas superiores. La cobertura de 2x permite obtener datos sobre más individuos que no a 5x o 7x, reduciendo el coste de genotipado masivo para cada individuo. Una cobertura de 1x permitiría aumentar el número de individuos a estudiar, pero se observó que las predicciones eran peores que a 2x. El motivo fue que a una cobertura de 1x (disponiendo de un promedio de una única lectura) es muy difícil distinguir los heterocigotos. En el trabajo de Ros-Freixedes et al., 2020b también se valora la cobertura de 2x como la óptima para realizar en los estudios poblacionales.

La capacidad de detección de SNPs es mucho mayor en la secuenciación (incluso a 1x) que la conseguida por medio de un chip de alta densidad (55K), convirtiendo a la secuenciación en una herramienta muy potente para detectar variantes, ya que permite detectar aquellas variantes que no se incluyeron en el diseño del chip.

Por otro lado, se determinó que genotipar los animales con más progenie con chips de genotipado de alta densidad no proporcionaba mejoras en la precisión de la imputación respecto a si los mismos animales se genotipaban a baja densidad. Se perdía información genotípica al reducirse el presupuesto destinado a chips de baja densidad y además con los chips de LD se obtenía suficiente información y por eso al aumentar el número de marcadores no se observaba una gran mejora en las concordancias.

Este test también demostró que genotipando la mitad de la población se reducía mucho la precisión, ya que, secuenciando el 50% de la población se disponían de la mitad de datos y por tanto no se podía realizar una imputación tan exitosa.

Otro parámetro que influye en los resultados es la relación genética entre individuos. En nuestro caso, al tratarse de una población con varias generaciones solapadas, la estimación de las probabilidades de segregación pudo ser más precisa y, por tanto, también la imputación de los genotipos faltantes.

La unión entre ambas metodologías (chip y secuenciación) parece ser un recurso de interés tanto económico como para la obtención de genotipos precisos.

5.3.2 Segundo test

Dividiremos los presupuestos en dos categorías; presupuestos bajos (150k a 450k), en los que no se dispone de suficiente capital para genotipar la totalidad de la población, y presupuestos altos (600k a 900k), donde se dispone de dinero suficiente para genotipar por chip todos los individuos.

Para un presupuesto bajo (150.000€, 300.000€, 450.000€), se determinó que es mejor genotipar el máximo número de animales por chip, ya que son más económicos y permiten obtener genotipos de más individuos, utilizados el primer paso de la imputación (estimación de las probabilidades de segregación), que se hace a partir de los genotipos obtenidos por chip. Sin embargo, como consecuencia, se destina poco presupuesto para secuenciar un número bajo de animales. En el caso de los presupuestos más bajos (150.000€ y 300.000€) <10 €/animal, la mayor parte de la inversión se destinó a los chips y poco a la secuenciación, por lo que las concordancias fueron bajas (< 85%). Con un presupuesto de 450.00€ la concordancia aumenta alrededor de un 5%. Con este capital parece que empieza a establecerse una tendencia entre el presupuesto y el porcentaje destinado para a cada método, de tal manera que se alcanza un pico alrededor del 80% de cerdos genotipados por chip, siendo este el porcentaje óptimo para entrelazar los resultados y crear haplotipos correctos, aumentando las concordancias.

Los resultados obtenidos para presupuestos superiores (600.000€, 750.000€, 900.000€) indican, incluso para el caso de menor capital invertido (600.000€), mayor precisión de imputación (aumentando un 5% la precisión al compararlo con 450.000€). Al poder invertir más en secuenciación, se observa que el porcentaje del capital destinado a los chips de genotipado va disminuyendo. Para 750.000€ se obtiene un pico alrededor del 60% y para 900.000€ en torno al 50% de los individuos genotipados por chip. Esto permite invertir más dinero en secuenciación y así obtener concordancias altas (> 95%).

Para presupuestos elevados aparece una tendencia a medida que se genotipan más cerdos mediante chip, pero a presupuestos menores no sigue esta tendencia. La lógica encontrada es que si todos los animales poseen datos de chip pero no hay suficientes datos de secuenciación, la imputación no funciona bien, ya que el segundo paso de la imputación se basa en la disponibilidad de datos de secuenciación. Lo mismo sucederá en la situación contraria: si todos tienen datos de secuenciación, pero no de chip, entonces el que falla es el primer paso de la imputación, que se basa en la disponibilidad

de datos de chip de genotipado. Esto hace que los presupuestos menores se beneficien más al invertir chips que en secuenciación, ya que disponer de suficientes datos sobre el genotipado en el primer paso de la imputación es el factor limitante para la precisión del segundo paso.

Se realizó un estudio paralelo donde se varió el porcentaje de capital inicial (utilizando los mismos presupuestos que en el test dos) y el porcentaje de animales genotipados en la población (10%, 20% …. 90% de 30.000) por medio de chips. A presupuestos menores, las mejores concordancias se encontraban al genotipar menos animales por medio de los chips, y al aumentar el presupuesto mayor era más cantidad de animales genotipados por medio del chip. Los presupuestos 150.000€, 300.000€, 450.000€, 600.000€, 750.000€, 900.000€ consiguieron el pico de animales genotipados por medio de chips respetivamente sobre el 20%, sobre el 50%, 80%, 100%, 100% de los 30.000 animales. Estos resultados nos mostraron un paralelismo con el segundo test. Observamos que a presupuestos bajos (150.000€, 300.000€, 450.000€) la mayor parte del presupuesto es destinada a chips, siendo respectivamente 90.000€, 225000€, 360.000€. Esto puede deberse a que el primer paso de imputación requiere datos de chip y a que el balance entre el número de datos recopilados por ambas técnicas se consigue al invertir más en chips (ya que estos detectan muchos menos SNP).

Para concluir en lo referente al estudio económico una cobertura de 2x es la óptima para estudios a nivel poblacional, al permitirnos analizar más animales. Si se necesita un estudio individual se tendría que realizar a coberturas superiores. Se conseguirán concordancias genotípicas mayores al genotipar la mayor cantidad de animales por chips de genotipado LD, si se utiliza el algoritmo de imputación a baja cobertura comentado. Dependiendo del presupuesto de partida y utilizando la herramienta de imputación comentada será más beneficioso invertir en chips (presupuestos pequeños) o en secuenciación (presupuestos elevados).

6. CONCLUSIONES

1. Por medio de un GWAS con los datos de 395 cerdos Duroc se detectó el gen *ATP1A1* como gen candidato asociado al pH de la carne.

2.1. El índice de detección de variantes varía según la cobertura. Con coberturas moderadas (≥5x) se detectaron más del 95% de variantes presentes en el chip, pero este porcentaje disminuyó a alrededor del 92% (83%) con una cobertura de 2x (1x).

2.2. Se consiguieron concordancias genotípicas y alélicas superiores al 80% a coberturas moderadas (≥5x), pero al disminuir la cobertura a 2x y 1x se redujo este porcentaje entorno al 58% y 35%, respectivamente.

2.3. Se validó la aplicación de secuenciación a coberturas moderadas (7x o 5x), como una tecnología viable para la identificación de variantes asociadas a caracteres productivos a nivel individual. Para la detección de caracteres productivos a nivel poblacional se identificaron las mejores concordancias a una cobertura de 2x cuando la aplicación de técnicas de imputación es viable. Con una cobertura de 1x la concordancia genotípica disminuía mucho porque no se conseguían detectar los heterocigotos.

3.1. Es importante disponer de datos de chip de genotipado del mayor número posible de animales de la población. La densidad de los chips usados (LD o HD) no tiene un efecto sobre la precisión de la imputación con algoritmos de imputación adecuados.

3.2. Con presupuestos limitados (<10 €/animal), es más beneficioso invertir en chips de genotipado que en secuenciación, ya que la disponibilidad de estos datos es el factor limitante para la precisión de la imputación. Solo cuando ya se dispone de un número elevado de animales genotipados con chip, es conveniente secuenciar, lo cual requiere mayores presupuestos.

7. REFERENCIAS

Ángel-Marín, P. A., Cardona-Cadavid, H., & Cerón-Muñoz, M. F. (2013). Genómica en la producción animal. *Revista Colombiana de Ciencia Animal-RECIA*, 497-518.

Busquets, X., & Agustí, A. G. N. (2001). Chip genético (ADN array): el futuro ya está aquí. *Archivos de Bronconeumología*, *37*(9), 394-396.

Cañón, J. (2006). Utilización de información molecular en programas de mejoramiento animal. *Ciencia y Tecnología Agropecuaria*, *7*(1), 5-15.

CM Dekkers, J. (2012). Application of genomics tools to animal breeding. *Current genomics*, *13*(3), 207-212.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), s13742-015.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., ... & Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, *14*(1), 128.

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., Van Binsbergen, R., Brøndum, R. F., ... & Esquerré, D. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature genetics*, *46*(8), 858.

Das, A., Panitz, F., Gregersen, V. R., Bendixen, C., & Holm, L. E. (2015). Deep sequencing of Danish Holstein dairy cattle for variant detection and insight into potential loss-of-function variants in protein coding genes. *BMC genomics*, *16*(1), 1043.

Davoli, R., Zappaterra, M., & Zambonelli, P. (2019). Genome-wide association study identifies markers associated with meat ultimate pH in Duroc pigs. *Animal genetics*, *50*(2), 154-156.

DePristo, MA, Banks, E., Poplin, R., Garimella, KV, Maguire, JR, Hartl, C., ... y McKenna, A. (2011). Un marco para el descubrimiento de variaciones y el genotipado utilizando datos de secuenciación de ADN de próxima generación. *Nature genetics* , *43* (5), 491.

Estany Illa, J., & Pena i Subirà, R. N. (2017). La selección genómica. *Suis, 2017, num. 140, p. 32-38.*

Faux, A. M., Gorjanc, G., Gaynor, R. C., Battagin, M., Edwards, S. M., Wilson, D. L., ... & Hickey, J. M. (2016). AlphaSim: software for breeding program simulation. *The plant genome*, *9*(3), 1-14.

Fujita, R. (2007). Genómica y su aplicación en producción animal.

Gomez-Sanchez, C. E., Kuppusamy, M., & Gomez-Sanchez, E. P. (2015). Somatic mutations of the ATP1A1 gene and aldosterone-producing adenomas. *Molecular and cellular endocrinology*, *408*, 213-219.

Gudbjartsson, D. F., Sulem, P., Helgason, H., Gylfason, A., Gudjonsson, S. A., Zink, F., ... & Sigurdsson, G. T. (2015). Sequence variants from whole genome sequencing a large group of Icelanders. *Scientific data*, *2*(1), 1-11.

Hickey, J. M., Gorjanc, G., Cleveland, M. A., Kranis, A., Jenko, J., & Mészáros, G. (2013). Sequencing millions of animals for genomic selection 2.0. *J Anim Breed Genet*, *130*(5), 331-332.

Iniesta, R., Guinó, E., & Moreno, V. (2005). Análisis estadístico de polimorfismos genéticos en estudios epidemiológicos. *Gaceta Sanitaria*, *19*(4), 333-341.

Jiang, Y., Jiang, Y., Wang, S., Zhang, Q., & Ding, X. (2019). Optimal sequencing depth design for whole genome re-sequencing in pigs. *BMC bioinformatics*, *20*(1), 556.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., ... & McDermott, M. G. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, *44*(W1), W90-W97.

Mordoh, A. (2019). Secuenciación masiva de ADN: la próxima generación. *Dermatología Argentina*, *25*(1), 02-08

Poplin, R., Ruano-Rubio, V., DePristo, MA, Fennell, TJ, Carneiro, MO, Van der Auwera, GA, ... & Shakir, K. (2017). Escalando el descubrimiento preciso de variantes genéticas a decenas de miles de muestras. *BioRxiv*

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), 559-575.

Ros-Freixedes, R., Reixach, J., Tor, M., & Estany, J. (2012). Expected genetic response for oleic acid content in pork. *Journal of animal science*, *90*(12), 4230-4238.

Ros-Freixedes, R., Gol, S., Pena, R. N., Tor, M., Ibanez-Escriche, N., Dekkers, J. C., & Estany, J. (2016). Genome-wide association study singles out SCD and LEPR as the two main loci influencing intramuscular fat content and fatty acid composition in Duroc pigs. *PLoS One*, *11*(3).

Ros-Freixedes, R., Gonen, S., Gorjanc, G., & Hickey, J. M. (2017). A method for allocating low-coverage sequencing resources by targeting haplotypes rather than individuals. *Genetics Selection Evolution*, *49*(1), 78.

Ros-Freixedes, R., Battagin, M., Johnsson, M., Gorjanc, G., Mileham, A. J., Rounsley, S. D., & Hickey, J. M. (2018). Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing. *Genetics Selection Evolution*, *50*(1), 64.

Ros-Freixedes, R., Whalen, A., Chen, C. Y., Gorjanc, G., Herring, W. O., Mileham, A. J., & Hickey, J. M. (2020a). Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations. *Genetics Selection Evolution*, *52*(1), 1-15.

Ros-Freixedes, R., Whalen, A., Gorjanc, G., Mileham, A. J., & Hickey, J. M. (2020b). Evaluation of sequencing strategies for whole-genome imputation with hybrid peeling. *Genetics Selection Evolution*, *52*(1), 1-19.

Sahoo, S. S., Mishra, C., Rout, M., Nayak, G., Mohanty, S. T., & Panigrahy, K. K. (2016). Comparative in silico and protein-protein interaction network analysis of ATP1A1 gene. *Gene Reports*, *5*, 134-139.

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, *19*(8), 491-504.

Schlingmann, K. P., Bandulik, S., Mammen, C., Tarailo-Graovac, M., Holm, R., Baumann, M., ... & Beck, B. B. (2018). Germline de novo mutations in ATP1A1 cause renal hypomagnesemia, refractory seizures, and intellectual disability. *The American Journal of Human Genetics*, *103*(5), 808-816.

Stindl, J., Tauber, P., Sterner, C., Tegtmeier, I., Warth, R., & Bandulik, S. (2015). Pathophysiology of Na+/K+-atpases in aldosterone secretion. *Experimental and Clinical Endocrinology & Diabetes*, *123*(03), P09_23.

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics*, *30*(9), 418-426.

Van Eenennaam, A. L., Weigel, K. A., Young, A. E., Cleveland, M. A., & Dekkers, J. C. (2014). Applied animal genomics: results from the field. *Annu. Rev. Anim. Biosci.*, *2*(1), 105-139.

Whalen, A., Ros-Freixedes, R., Wilson, D. L., Gorjanc, G., & Hickey, J. M. (2018). Hybrid peeling for fast and accurate calling, phasing, and imputation with sequence data of any coverage in pedigrees. *Genetics Selection Evolution*, *50*(1), 1-15.

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P. C., Hu, L., ... & Matsuoka, M. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature genetics*, *48*(8), 927.

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., ... & Bhai, J. (2020). Ensembl 2020. *Nucleic acids research*, *48*(D1), D682-D688.